Bayesian Multivariate Regression Analysis with a New Class of Skewed Distributions

José T.A.S. Ferreira and Mark F.J. Steel*

Department of Statistics
University of Warwick, UK

Abstract

In this paper, we introduce a novel class of skewed multivariate distributions and, more generally, a method of building such a class on the basis of univariate skewed distributions. The method is based on a general linear transformation of a multidimensional random variable with independent components, each with a skewed distribution. Our proposed class of multivariate skewed distributions has a simple, intuitive form for the pdf, moment existence only depends on the existence of the moments of the underlying symmetric univariate distributions, and we avoid any conditioning on unobserved variables. In addition, we can freely allow for any mean and covariance structure in combination with any magnitude and direction of skewness. In order to deal with both skewness and fat tails, we introduce multivariate skewed regression models with fat tails, based on Student distributions. We present two main classes of such distributions, one of which is novel even under symmetry. Under standard non-informative priors on both regression and scale parameters, we derive conditions for propriety of the posterior and for existence of posterior moments. We describe MCMC samplers for conducting Bayesian inference and analyse two applications, one concerning the distribution of various measures of firm size and another on a set of biomedical data.

Keywords: Asymmetric distributions; Heavy tails; Linear regression model; Mardia's measure of skewness; Orthogonal matrices; Posterior propriety.

1 Introduction

In recent years there has been an increasing interest in more flexible distributions that can represent observed behaviour more closely.

The contribution of this article is twofold. We start by introducing a general method for the definition of multivariate skewed distributions. Then we use this new class of distributions in a multivariate regression context and propose Bayesian inference procedures.

The class of elliptical distributions, as presented *e.g.* by Kelker (1970), has been the predominant framework for multivariate continuous random quantities. This class of distributions is quite well studied and we refer the interested reader to *e.g.* Fang et al. (1990). However, in a substantial number

*J. Tomé Ferreira is Ph.D. student (email: J.T.A.Ferreira@warwick.ac.uk) and Mark Steel is Professor (email M.F.Steel@stats.warwick.ac.uk), Department of Statistics, University of Warwick, Coventry, CV4 7AL, U.K. The work of Ferreira was supported by grant SFRH BD 1399 2000 from Fundação para a Ciência e Tecnologia, Ministério para a Ciência e Tecnologia, Portugal. The UK Data Archive bears no responsibility for the analysis or interpretation contained in this article.

of situations, elliptical distributions have been found to be too restrictive. Such is certainly the case for problems where the random quantity exhibits skewness, our main focus in the present article.

So far, the literature on skewed distributions has mainly dealt with univariate cases. Azzalini and Dalla Valle (1996) is one of the first multivariate proposals. Based on the univariate skew-Normal distribution analysed in detail by Azzalini (1985), this method can be interpreted as defining a multivariate skew-Normal density by conditioning on an unobserved argument. Such conditioning models, also known as hidden truncation models (Arnold and Beaver 2000), have been generalised further. Still conditioning on one unobserved variable, Branco and Dey (2001) introduced a class of multivariate skew-elliptical distributions, and Arnold and Beaver (2002) made these models more general by allowing for non-elliptical skew distributions. Within the class of hidden truncation models, but conditioning on as many arguments as observed variables, Sahu et al. (2003) generated a very general class of multivariate skew-elliptical distributions. A recent review of the literature on skewed distributions generated by hidden truncation models can be found in Arnold and Beaver (2002).

A different approach to multivariate skewed distributions was proposed by Jones (2002). Starting with spherically symmetric distributions, the author proposed replacing the marginal distribution of some of the variables by a skewed distribution. The method is particularly interesting when only one variable is to have a skewed marginal, as several options for univariate skewed distributions are available in the literature.

The class that we introduce in this article is based on a general linear transformation of a multidimensional random variable with independent components, each having a skewed distribution, with probability density function (pdf) constructed using the method introduced in Fernández and Steel (1998). This method of constructing a multivariate distribution on the basis of a univariate one does not require any additional restrictions beyond the ones imposed on univariate distributions in Fernández and Steel (1998). There, the authors present a method to transform any symmetric, unimodal distribution into a skewed distribution. Our proposal for multivariate skewed distributions has the advantages that the pdf has a simple, intuitive form, moment existence is only dependent on the existence of the moments of the underlying symmetric univariate distributions, and we avoid any conditioning on unobserved variables. In addition, we can freely allow for any mean and covariance structure in combination with any magnitude and direction of skewness.

Despite focusing on this class, we highlight that it is possible to use any other general method for generating univariate skewed distributions for the independent components. For example, we could base ourselves on the univariate distributions introduced in Azzalini (1985), Azzalini and Capitanio (2003) or Jones and Faddy (2003).

A proposal for multivariate skewed distributions using a linear combination of independent univariate skewed distributions has appeared before in Bauwens and Laurent (2002). However, the one we present here is fundamentally different, as will be explained in the sequel. Hoggart, Walker and Smith (2003) use an orthogonal transformation of random variables with symmetric univariate distributions to create a bivariate distribution with different kurtosis in each direction, which, however, does not allow for skewness.

Subsequently, we introduce multivariate skewed regression models with fat tails, by considering a linear regression structure with skewed and heavy-tailed error terms. In order to allow for heavy tails we use skewed versions of Student-t distributions. We consider standard non-informative priors on both regression and scale parameters. Skewness and tail behaviour are not fixed but inferred from the data. We derive conditions that make Bayesian analysis feasible (*i.e.* lead to a proper posterior), under the improper prior structure. In addition, we provide results on the existence of posterior moments of the regression coefficients and the determinant of the scale matrix.

We introduce two different Student-based multivariate regression models. One can be represented as a scale mixture of multivariate Normals, and is, thus, characterized by one single mixing variable. Therefore, this leads to the skewed analogue of the multivariate Student-t regression model in Fernández and Steel (1999). In the latter symmetric model, there is no need to use the orthogonal transformations that we introduce in this paper, since the model is based on a multivariate Normal, which is spherical. The moment we introduce skewness such an orthogonal transformation becomes crucial as a means of specifying the directions of the skewness. The other class of heavy-tailed models that we introduce here is based on a transformation of independent Student-t distributed random variables. As this class of distributions is no longer based on a spherical class, we need to use the orthogonal transformations introduced in the sequel, even under symmetry. Thus, the present paper also introduces an, as yet unexplored, class of symmetric heavy-tailed distributions and sheds light on its properties regarding Bayesian inference.

Inference in our regression setup is performed using hybrid Markov chain Monte Carlo (MCMC) samplers using data augmentation. With current computational power, inference can easily be performed even for relatively large problems.

We illustrate the flexibility of the proposed framework by an application to the size distribution of a group of small and medium British exporting firms, and in a regression problem using biomedical data from the Australian Institute of Sport.

Section 2 briefly recalls the univariate skewed distributions of Fernández and Steel (1998), introduces the multivariate skewed distributions, together with some properties, provides a useful parameterisation of these distributions and presents key examples. Section 3 develops the Bayesian multivariate skewed regression models, studies the effect of skewness on the existence of posterior moments, and assesses the feasibility of inference under asymmetric, heavy-tailed sampling. Section 4 is devoted to the numerical implementation employed to conduct inference. In Section 5 we present the applications. Finally, Section 6 provides some concluding remarks. All proofs are deferred to the Appendix, without explicit mention in the body of the text.

2 Skewed Distributions

2.1 The univariate case

Fernández and Steel (1998) propose a method for introducing skewness into a unimodal distribution symmetric around the origin. The basic idea is to introduce inverse scale factors in the positive and the negative half real lines. Let $f(\cdot)$ be a univariate pdf that is symmetric around zero, and such that $f(s)$ is assumed to be decreasing in the absolute value of s. Also, let γ be a scalar in $(0, \infty)$. Then, the skewed distribution on the real line is given by the pdf

$$p(\epsilon|\gamma, f) = \frac{2}{\gamma + \frac{1}{\gamma}} \left\{ f\left(\frac{\epsilon}{\gamma}\right) I_{[0,\infty)}(\epsilon) + f(\gamma\epsilon) I_{(-\infty,0)}(\epsilon) \right\} = \frac{2}{\gamma + \frac{1}{\gamma}} f\left(\epsilon \gamma^{-\text{sign}(\epsilon)}\right), \tag{1}$$

where $I_S(\cdot)$ is the indicator function on S, and sign$(\cdot)$ is the usual sign function in $\Re$.

There are several interesting characteristics of the skewed density given in (1). If the skewness parameter γ is unity, then we retrieve the original symmetric density. The mode of the density is unchanged, remaining at zero irrespective of the particular value of γ. Also, the probability mass assigned to each side of the mode is independent of $f(\cdot)$ and given by

$$P(\epsilon > 0 \,|\, \gamma, f) = \frac{\gamma^2}{1+\gamma^2},$$

allowing γ to parameterise the complete range of mass on each side of the origin. Another relevant feature of this method is that the existence of moments of $p(\epsilon|\gamma, f)$ does not depend on γ, but only on the existence of moments of the initial, symmetric density $f(\cdot)$. Furthermore, the moments can be written simply as functions of the moments of $f(\cdot)$. The rth moment ($r \in \Re$) is obtained as

$$E(\epsilon^r|\gamma, f) = M_r \frac{\gamma^{r+1} + \frac{(-1)^r}{\gamma^{r+1}}}{\gamma + \frac{1}{\gamma}} \quad \text{where} \quad M_r(f) = \int_0^\infty s^r 2f(s)ds. \tag{2}$$

Finally, simple manipulation reveals that $p(\epsilon|\gamma, f) = p(-\epsilon|1/\gamma, f)$.

2.2 The multivariate case

2.2.1 Definition

The construction of multivariate skewed distributions presented here is based on linear transformations of univariate skewed distributions. Let m be the dimension of the random variable $\epsilon = (\epsilon_1, \ldots, \epsilon_m)' \in \Re^m$ and $\gamma = (\gamma_1, \ldots, \gamma_m)' \in \Re^m_+$. Further, let $f = (f_1(\cdot), \ldots, f_m(\cdot))'$ denote a vector of m unimodal and symmetric univariate pdfs. The pdf of the multivariate skewed distribution with independent components is given by

$$p(\epsilon|\gamma, f) = \prod_{j=1}^m p(\epsilon_j|\gamma_j, f_j), \tag{3}$$

where each $p(\epsilon_j|\gamma_j, f_j)$ is as in (1).

Following an affine linear transformation, given a vector $\mu = (\mu_1, \ldots, \mu_m)'$ and a non-singular matrix $A \in R^{m\times m}$, the variable $\eta = (\eta_1, \ldots, \eta_m)' \in R^m$, defined as

$$\eta = A'\epsilon + \mu \tag{4}$$

has a general multivariate skewed distribution, with parameters $\mu,\ A, \gamma$ and f, denoted by $Sk_m(\mu, A, \gamma, f)$. The pdf for η is given by

$$p(\eta|\mu, A, \gamma, f) = \|A\|^{-1} \prod_{j=1}^m p[(\eta-\mu)' A^{-1}_{\cdot j}|\gamma_j, f_j], \tag{5}$$

where $A^{-1}_{\cdot j}$ denotes the j-th column of A^{-1}, $\|A\|$ denotes the absolute value of the determinant of A, and $p(\cdot|\gamma_j, f_j)$ is as in (1). The distribution of η is unimodal with mode μ, A introduces the dependence between the components of η, while γ determines the skewness of the independent components of ϵ.

Figure 1 presents contour plots for four different bivariate skewed distributions, with both $f_1(\cdot)$ and $f_2(\cdot)$ equal to $\phi(\cdot)$, the univariate standard Normal pdf, and μ set to the zero vector. Figure 1 (a) represents the density of a distribution with independent components, where only one of these is (positively) skewed. The remaining three plots were all obtained using the same values for the skewness components, namely $\gamma = (0.5, 1.5)'$. By varying the transformation matrix A it is possible to obtain a diverse set of shapes for the density. If A equals the identity matrix, then the effect of γ

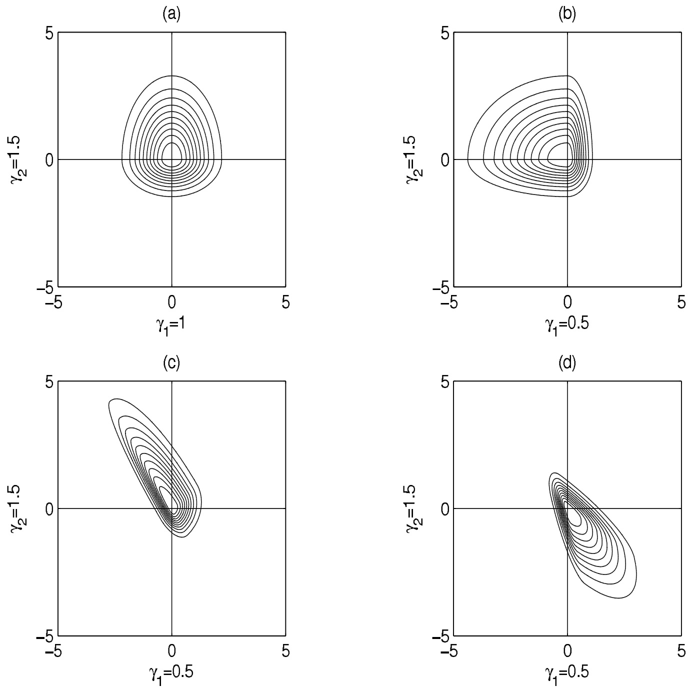

Figure 1: Contour plots of four different bivariate distributions. Plots (a) and (b) correspond to $A = I$, whereas A for plots (c) and (d) was chosen to lead to the same $A'A$ matrix.

is evident (see Figure 1 (b)). In the context of skewed distributions with independent components, γ values larger than one always correspond to a positively skewed marginal, and the reverse happens for values of $\gamma \in (0, 1)$. Figures 1 (c) and (d) represent skewed distributions with dependent components. It can be seen that the shape of the contours varies extensively, even with the same γ, highlighting the flexibility of the method we introduce. To further illustrate the role of the matrix A, we have generated plots (c) and (d) with the same matrix $A'A = [\frac{1}{2} \quad -\frac{1}{2}; -\frac{1}{2} \quad 1]$. As discussed in Subsection 2.2.3, $A'A$ is all that would matter without skewness.

2.2.2 Moments

Calculation of the moments of η is straightforward and is achieved using the moments of the, much simpler, univariate pdfs $f_j(\cdot),\ j = 1, \ldots, m$. Further, like in the univariate case, the existence of the moments of η depends only on $f_j(\cdot)$ and not on the skewness parameters. Due to the linear transformation used in (4), the existence of the rth positive moment of η depends exclusively on the existence of the first r moments of the distributions with density $f_j(\cdot)$. As an illustration, assuming a common $f_j(\cdot) = f(\cdot),\ j = 1, \ldots, m$, the mean vector and the covariance matrix of η are given by

$$E(\eta) = \mu + M_1 A' \begin{pmatrix} \gamma_1 - \frac{1}{\gamma_1} \\ \cdots \\ \gamma_m - \frac{1}{\gamma_m} \end{pmatrix},$$

and

$$Var(\eta) = A' \left\{ Diag \left[\left(M_2 - M_1^2 \right) \left(\gamma_j^2 + \frac{1}{\gamma_j^2} \right) + 2M_1^2 - M_2 \right]_{j=1,\ldots,m} \right\} A,$$

as long as M_1 and M_2, given by (2), both exist.

Thus, even though $E(\eta)$ and $Var(\eta)$ depend on γ directly, their values are not restricted by it. We can obtain any desired mean and covariance values for the distribution even after setting γ, simply by choosing μ and A appropriately. We feel this is an advantage of this class of skewed distributions, when compared to proposals such as the ones introduced by Azzalini and Dalla Valle (1996) or Sahu et al. (2003). In these, the set of covariances obtainable after setting the parameters regulating the skewness of the distribution is restricted. In contrast, our framework allows for independent modelling of mean, covariance and skewness.

Figure 2 illustrates how we can fix the covariance and generate quite different distributions by changing both A and γ. All the contour plots in Figure 2 represent distributions with identity covariance matrix, but with quite different shapes.

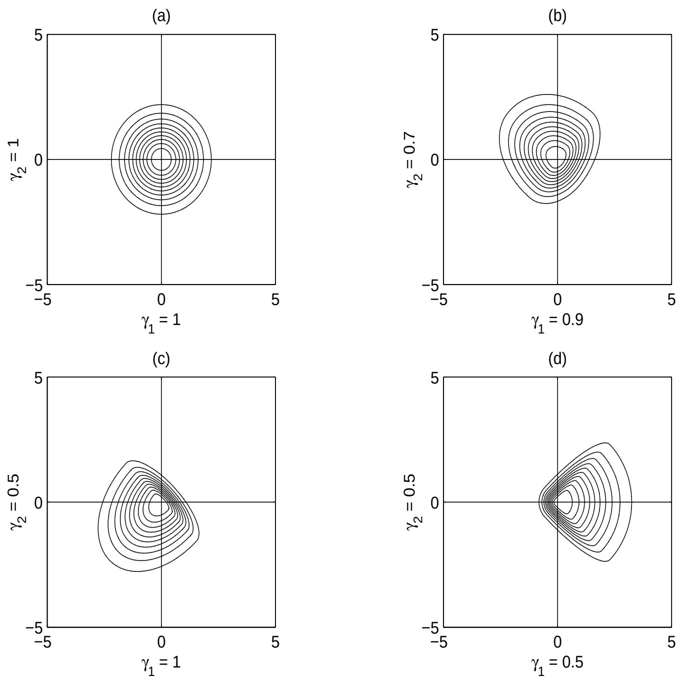

Figure 2: Contour plots of four pdfs with identical covariance matrix (equal to the identity matrix).

This ability of our class of skewed distributions to cover all possible mean and covariance structures in linked with one potential drawback, and that is the fact that the class of distributions is not closed under marginalisation. This results from the fact that a linear combination of random quantities with pdf as in (1) does not necessarily have a density of the same form. In a bivariate context, Moran (1967) remarks that a necessary condition for classes of distributions with fixed marginals to cover the entire range of values for the correlation coefficient is that the marginals are symmetric.

Not many measures of multivariate skewness have been proposed in the literature. One measure of multivariate skewness is $\beta_{1,m}$, introduced by Mardia (1970) and given by

$$\beta_{1,m}(\eta) = \sum_{r,s,t}^{m} \sum_{r',s',t'}^{m} \sigma^{rr'}\sigma^{ss'}\sigma^{tt'} E\left[(\eta_r - \alpha_r)(\eta_s - \alpha_s)(\eta_t - \alpha_t)\right] E\left[(\eta_{r'} - \alpha_{r'})(\eta_{s'} - \alpha_{s'})(\eta_{t'} - \alpha_{t'})\right],$$

where α_j and $\sigma^{jj'}$, $j,\ j' = 1, \ldots, m$ denote the elements of the mean vector and precision matrix of η, respectively. Two main characteristics of $\beta_{1,m}$ make it interesting for use: it equals zero for any symmetric distribution, with unimodal asymmetric distributions being characterised by values of the measure larger than zero, and it is invariant under non-singular affine transformations.

As $\beta_{1,m}$ is invariant under non-singular affine transformations, the calculation of its value for a multivariate skewed distribution generated using the construction we propose is trivial. Let $\eta \sim Sk_m(\mu, A, \gamma, f)$, then by making use of an alternative affine transformation of the original variables ϵ it is possible to obtain a set of variables $\psi \sim Sk_m(\mu^*, A^*, \gamma, f)$, with A^* diagonal, such that $Var(\psi)$ equals the identity matrix and $E(\psi)$ is zero. Now, as A^* is diagonal, by (5), the components of ψ are independent and Mardia's skewness measure is given by

$$\beta_{1,m}(\eta) = \beta_{1,m}(\psi) = \sum_{j=1}^{m} \left[E\left(\psi_j^3\right)\right]^2, \tag{6}$$

which, from (2), is straightforward to calculate and does not depend on μ or A. This ease of calculating Mardia's measure of skewness, for any pdfs $f_j(\cdot)$, is not shared by any of the methods based on conditioning on unobserved arguments or marginal replacement. The expression in (6) also shows that each particular γ_j has a contribution to the measure that is independent of the remaining elements of γ. As a consequence, if γ_j is set to one, its contribution to (6) vanishes. The existence of $\beta_{1,m}$ depends exclusively on the existence of $M_3(f_j),\ j = 1, \ldots, m$ defined in (2).

Figure 3 plots $\beta_{1,2}$ as a function of $\gamma \in (0,1] \times (0,1]$ with $f_j(\cdot) = \phi(\cdot),\ j = 1, 2$. As expected, $\beta_{1,2}$ is a continuous, strictly decreasing function of γ_j in $(0,1]$, $j = 1, 2$. Other values of γ are covered by the fact that the value of $\beta_{1,2}$ is unaffected by inverting either γ_1, γ_2 or both. The value of $\beta_{1,2}$ is bounded below by zero (symmetric case) and

$$\lim_{\gamma \to 0} \beta_{1,2} = 4\frac{(4-\pi)^2}{(\pi-2)^3} \approx 1.96. \tag{7}$$

These same bounds are obtained in Sahu et al. (2003) for the authors' definition of the skew-Normal distribution. For the skew-Normal distribution of Azzalini and Dalla Valle (1996), the upper bound is half the value in (7).

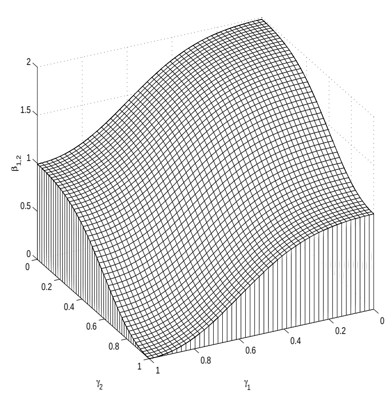

Figure 3: Plot of $\beta_{1,2}$ for a bivariate skew-Normal distribution as a function of γ.

2.2.3 The importance of orthogonal transformations

In the sequel, we make use of the following result on the decomposition of nonsingular matrices.

Lemma 1. If A is any $m \times m$ real non-singular matrix, there exists an orthogonal matrix O_U such that $A = O_U U$, where U is a real upper triangular matrix with positive diagonal elements. Likewise, there exists another orthogonal matrix O_L such that $A = LO_L$, where L is a real lower triangular matrix with positive diagonal elements. Both representations are unique.

In order to gain further insight into the full effect of A, suppose that $A = LO$ (as defined in Lemma 1) and, for simplicity assume that $\mu = 0$. From (4) we then have $\eta = O'L'\epsilon$, indicating that ϵ is first subjected to a linear transformation, and then to a rotation if $|O| = 1$ or a rotoinversion if $|O| = -1$. If O is the identity matrix, the j-th component of η is a linear combination of the last $m - j + 1$ components of ϵ. The effect of O is that it rotates and/or reflects the axes along which the joint distribution is a linear combination of the last $m - j + 1$ components of ϵ. Figure 4 exemplifies the effect of O, using bivariate skewed distributions. In Figure 4 (a) $A = LO_a$, while in Figure 4 (b) $A = LO_b$, with μ equal to zero and

$$L = \begin{pmatrix} 1 & 0 \\ \frac{1}{4} & 1 \end{pmatrix}, O_a = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, O_b = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix} \text{ and } \gamma = \begin{pmatrix} \frac{3}{4} \\ \frac{3}{2} \end{pmatrix}.$$

Along the axes e_1 the distribution is given as a linear combination of two independent univariate skewed distributions with skewness parameters 3/4 and 3/2. Similarly, along the axes e_2 the distribution is a univariate skewed distribution with skewness parameter equal to 3/2. The contours in Figure 4 (b) can be obtained from the ones in Figure 4 (a) by reflecting them about any of the axes and rotating them. Inspired by the representation $A = LO$, we define the *basic axis* e_j, $j = 1, ...m$ as the axis along

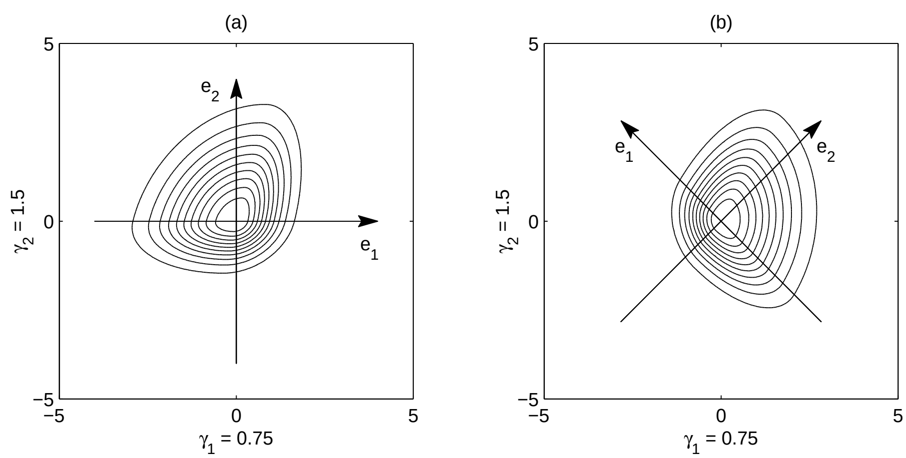

Figure 4: Contour plots of two bivariate skewed pdfs, together with their basic axes.

which the distribution is a linear combination of $m-j+1$ independent univariate skewed distributions with skewness parameters γ_k, $k = j, \dots, m$. Changing the orthogonal matrix O is then equivalent to performing a rotation of the basic axes, possibly after performing a reflection about some of them. As is evident from Figure 4 these axes define the direction of the skewness of the distribution.

A well-known fact from the theory of multivariate distributions is that if η is given by (4) and the distribution of ϵ belongs to the spherical class (*e.g.* the Normal distribution), then A only needs to be known up to $\Sigma = A'A$ or, equivalently, A can be either upper or lower triangular. Sphericity can be defined as distributional invariance with respect to orthogonal transformations (see *e.g.* Fang et al., 1990, p. 27), so that $\epsilon \overset{d}{=} O\epsilon$ for any orthogonal matrix O under spherical ϵ. It is then obvious that for $A = OU$ the transformed variable η in (4) has the same distribution as $U'\epsilon + \mu$, so that the choice of O is irrelevant. Equivalently, continuous spherical distributions on ϵ are characterised by a pdf that only depends on $\epsilon'\epsilon$, so it is clear that the induced pdf on η will only depend on $(\eta - \mu)'\Sigma^{-1}(\eta - \mu)$ and only $\Sigma = A'A = U'U$ matters. However, if ϵ is outside the spherical class then knowledge of Σ alone is no longer sufficient. For example, this is the case when the components of ϵ have independent

Student distributions or if their distributions are skewed. Thus, the orthogonal matrix O plays an important role. Not taking O into account (*e.g.*, by implicitly taking $O = I$) in skewed cases would imply favouring specific directions for the asymmetry of the distribution. In our view, it is essential for defining a general class of skewed distributions to also introduce parameters that specify the direction of the skewness, in our case by specifying the basic axes through A.

The skewed distribution of Sahu et al. (2003) introduces skewness into symmetric distributions along the coordinate axes. Bauwens and Laurent (2002) use a regression framework with a linear transformation as in (4), where ϵ has a similar distribution as in (3), but fix $A = \Sigma^{1/2}$, the spectral decomposition of Σ. The latter formulation does not allow for a separate choice of the directions of the asymmetry of the distribution, and fixes it to be a function of Σ.

2.3 Unique parameterisation of skewed distributions

The final part of the previous subsection shows that defining A via Σ is no longer sufficient for the class of distributions that we introduce in this article. Here we provide a unique parameterisation of our skewed distributions. However, even if the components of γ are set to unity, the parameterisation is still suitable, *i.e.* is unique, if the distribution of ϵ in (4) is not spherical.

Let $\eta \sim Sk_m(\mu, A, \gamma, f)$. As mentioned previously, the restrictions on parameters A and γ are that A is a non-singular matrix, and that γ is restricted to the positive orthant of the real space of appropriate dimension. However, this parameter space is not fully adequate in the sense that the same distribution of η can be defined using different sets of parameter values, which is an undesirable feature, especially for inference.

Let $r = (r_1, \ldots, r_m)'$ be a permutation of the first m positive integers, A_r be the $m \times m$ matrix where the jth row is the r_jth row of A, and γ_r be the m-dimensional vector where the jth element is the r_jth element of γ and define f_r similarly. Then, it follows directly from (4) that $Sk_m(\mu, A, \gamma, f) = Sk_m(\mu, A_r, \gamma_r, f_r)$. There are $m!$ different permutations r.

Also, let $s = (s_1, \ldots, s_m)'$ be a vector whose components are in $\{-1, 1\}$, A_s be the $m \times m$ matrix where the jth row equals the jth row of A times s_j, and let γ_s be the m-dimensional vector where component j is given by the jth element of γ to the power s_j. Then, it follows directly from the property of the univariate skew distributions stated at the end of Subsection 2.1 that $Sk_m(\mu, A, \gamma, f) = Sk_m(\mu, A_s, \gamma_s, f)$. The number of different vectors s is 2^m.

Combining both transformations gives all the $m!2^m$ parameter values that define the same distribution. These values are distinct if the components of γ are all distinct and different from unity.

There are several alternatives for reducing the parameter space in order to achieve a one-to-one parameterisation of the class of skewed distributions. Here we present one that is valid except for a set of distributions that, under most probability measures, will have zero mass. Making use of Lemma 1, through $A = OU$, we first reparameterise from (μ, A, γ, f) to (μ, O, U, γ, f) where O is an orthogonal matrix and U an upper triangular matrix with strictly positive diagonal elements. We can now create a one-to-one parameterisation by restricting the matrix $O = (O_{ij}),\ i, j = 1, \ldots, m$ to have

- $O_{11} > -O_{m1} > -O_{(m-1)1} > \ldots > |O_{21}| > 0$
- $|O| = (-1)^{m+1}$,

and by adjusting γ and f accordingly. The set of all such matrices O will be denoted by $\mathcal{O}^m$. This set of restrictions provides a one-to-one parameterisation, for all distributions with distinct components of

γ and matrices A without zeros in the first column. Indeed, if A_1 and A_2 differ by a signed permutation of rows (*i.e.* are equivalent), then $A_1 = O_1 U$ and $A_2 = O_2 U$, where O_1 and O_2 differ by the same signed permutation of rows. The two conditions above ensure that one and only one of such signed permutations is allowed in the parameter space.

2.4 Examples of multivariate skewed distributions

Even though it is possible to use symmetric, unimodal pdfs $f_j(\cdot),\ j = 1, \ldots, m$ from different parametric families, in what follows we will mainly focus on cases where for any $j = 1, \ldots, m$, $f_j(\cdot)$ can be written as $f_{\nu_j}(\cdot)$, with ν_j in some set $\mathcal{N}$. We then identify the multivariate skewed distribution generated by (5) with the name of the multivariate distribution that would result if $\gamma_j = 1,\ j = 1, \ldots, m$.

2.4.1 Skew-Normal

The multivariate skew-Normal distribution is obtained when $f_{\nu_j}(\cdot) = \phi(\cdot),\ j = 1, \ldots, m$, *i.e.* the pdf of the univariate standard Normal distribution. In this case, $\nu_j,\ j = 1, \ldots, m$, is vacuous.

2.4.2 Skew-Independent Student

The multivariate skew-Independent Student (skew-IStudent) with degrees of freedom (df) vector $\nu = (\nu_1, \ldots, \nu_m)'$ is generated when $f_{\nu_j}(\cdot)$ is the univariate Student pdf with $\nu_j \in \Re_+$ df, given by

$$f_{\nu_j}(x) = \frac{\Gamma\left(\frac{\nu_j+1}{2}\right)}{\Gamma\left(\frac{\nu_j}{2}\right)(\pi\nu_j)^{1/2}} \left[1 + \frac{x^2}{\nu_j}\right]^{-\frac{\nu_j+1}{2}} \tag{8}$$

2.4.3 Skewed mixture of Normals

Mixtures of Normals are an important class of distributions. Using a slight extension of the framework in (4), scale mixtures of Normals can be created by

$$\eta = \lambda^{-\frac{1}{2}} A'\epsilon + \mu \tag{9}$$

where ϵ follows a multivariate standard Normal distribution and a mixing distribution is assigned to λ. Skewed mixtures of Normals are defined in a similar way, by taking ϵ as in (3), with $f_j(\cdot) = \phi(\cdot),\ j = 1, \ldots, m$.

A particular case that will be used in the sequel is the skew-Student distribution with ν^* df, obtained if λ has a Gamma distribution with both shape and precision parameter set to $\nu^*/2$.

3 Regression Modelling

In the remainder we assume that we have n observations from an underlying process, given by pairs $(x_i, y_i),\ i = 1, \ldots, n$, where $x_i \in \Re^k$ is a vector of explanatory variables and $y_i \in \Re^m$ is the variable of interest. Throughout, we condition on x_i without mentioning it explicitly. The n observations are grouped in $X \in \Re^{n\times k}$, the design matrix, and $Y \in \Re^{n\times m}$, with each row corresponding to one observation.

Let us assume the observables $y_i \in \Re^m,\ i = 1, \ldots, n$, are generated from

$$y_i = g_i(B) + \lambda_i^{-\frac{1}{2}} A'\epsilon_i \tag{10}$$

where $g_i(\cdot)$ is a known measurable function in $\Re^m$, B parameterises the location, $A = OU$ is the transformation matrix for y_i, with $U \in \mathcal{U}^m$, the set of upper triangular $m \times m$ matrices with positive diagonal elements and $O \in \mathcal{O}^m$, the set of $m \times m$ orthogonal matrices that satisfy the conditions at the end of Subsection 2.3. $\lambda_i,\ i = 1, \ldots, n$ are independently drawn from some common underlying distribution P_λ on $\mathcal{L}$. We assume $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{im})',\ i = 1, \ldots, n$ to be independent and identically distributed conditionally on parameters $\nu \in \mathcal{N}^m$, and $\gamma = (\gamma_1, \ldots \gamma_m)'$ in $\Re^m_+$, with pdf as in (3).

3.1 Existence of moments under improper priors

We now consider the impact of introducing skewness into the multivariate sampling distribution on the existence of the posterior distribution and moments in the context of this general regression model.

Let $\lambda = (\lambda_1, \ldots, \lambda_n)'$. We adopt the following prior product structure:

$$P_{B,O,U,\lambda,\gamma,\nu} = P_{B,O,U} \times P_\lambda \times P_\gamma \times P_\nu. \tag{11}$$

The usual non-informative prior for regression modelling with elliptically distributed errors, is an improper prior on B and $\Sigma = A'A$ given by

$$p(B, \Sigma) = p(B)p(\Sigma) \propto |\Sigma|^{-\frac{m+1}{2}}. \tag{12}$$

We define a non-informative prior on B, O and U that is compatible with (12). From (12) and transforming from $\Sigma = A'A = U'U$ to U we have that $p(\Sigma) \propto |\Sigma|^{-\frac{m+1}{2}} \Leftrightarrow p(U) \propto \prod_{j=1}^m u_{jj}^{m-j} |U|^{-m}$. In addition, we take $p(O)$ such that its distribution on $\mathcal{O}^m$ is invariant to linear orthogonal transformations (see (32)-(33) and Appendix B.4). The prior on B is as in (12). Finally, we assume that P_λ and P_γ and P_ν are all proper distributions on $\mathcal{L}^n$, $\Re^m_+$ and $\mathcal{N}^m$, respectively. The full prior distribution is given by (11) with $P_{B,O,U}$ corresponding to

$$p(B, O, U) \propto p(O) \prod_{j=1}^{m} u_{jj}^{m-j} |U|^{-m}. \tag{13}$$

We can then derive the following result:

Theorem 1. Consider n independent replications from the sampling distribution given in (10) and the prior in (11) and (13). Denoting the lth element of B by B_l, $l = 1, \ldots, p$, and given $r_1, \ldots, r_p$ and $r \geq 0$, we obtain that for any P_γ

$$E\left(|\Sigma|^{r/2} \prod_{l=1}^{p} |B_l|^{r_l} \,\middle|\, Y \right) < \infty,$$

if and only if the same holds for inference with symmetrically distributed disturbances.

The result in Theorem 1 states that the existence of posterior moments of B and of non-negative posteriors moment of $|\Sigma|$ is unaffected by the extra vector of unknowns γ under any proper prior P_γ. Propriety of the posterior distribution is therefore not influenced by incorporating skewness in the sampling, as can be assessed by setting $r = r_1 = \cdots = r_p = 0$. This result extends Theorem 1 in Fernández and Steel (1998) to the case of multivariate skewed distributions.

We now completely specify two Bayesian models that account for both skewness and fat tails. Further, we provide results on posterior inference with these models. We define $g_i(B) = B'x_i$, where

$B \in \Re^{m\times k}$ (so, $p = mk$, and we shall now denote the elements of B by $B_{lj}, l = 1,\ldots,k, j = 1,\ldots,m$). This corresponds to the commonly used linear regression model. The complete design matrix $X = (x_1,\ldots,x_n)'$ will always be assumed to be of full rank, implying that $n \geq k$.

3.2 Inference under skew-Student sampling

The first of the models that we introduce here is the linear regression model, assuming that the errors have a skew-Student distribution, defined in Subsection 2.4.3. In particular we consider the following special case of the model in (10), (11) and (13):

- $f_{\nu_j}(\cdot) = \phi(\cdot),\ j = 1,\ldots,m.$
- For $i = 1,\ldots,n$, λ_i, given a positive parameter $\nu^* \in \mathcal{N}^*$, has a Gamma distribution with both parameters equal to $\nu^*/2$. The prior distribution on ν^*, P_{ν^*}, is proper.

Thus, we assume n independent replications of the sampling density

$$p(y_i|B,O,U,\nu^*,\gamma) = |U|^{-1}\prod_{j=1}^{m}\frac{2}{\gamma_j+\frac{1}{\gamma_j}}\int_{\Re_+}\lambda_i^{\frac{m}{2}}\ \phi\left(\lambda_i^{\frac{1}{2}}d_{ij}\ \gamma_j^{-\mathrm{sign}(d_{ij})}\right)p_G\left(\lambda_i\middle|\frac{\nu^*}{2},\frac{\nu^*}{2}\right)d\lambda_i, \tag{14}$$

with $d_{ij} = [O(U')^{-1}]_{j\cdot}(y_i - B'x_i)$.

The sampling distribution given in (14) will be denoted as m-dimensional skew-Student with location $B'x_i$, transformation OU, skewness parameter γ and ν^* df. Matrix B is usually of primary interest as it represents the regression coefficients. Also of common practical importance will be $\Sigma = U'U$ as it contains information about the dispersion of y. The remaining parameters have a well-defined purpose. Skewness is controlled jointly by γ and OU, while $\nu^* \in \Re_+$ determines the thickness of the tails of the multivariate distribution.

The results provided in this subsection will again extend results from Fernández and Steel (1998) to the multivariate case.

We begin by assessing the propriety of the posterior distribution.

Theorem 2. Consider n independent replications from the sampling model in (14) under the prior in (11) and (13). Then the posterior distribution is proper if and only if $n \geq m + k$, for any choices of P_{ν^*} and P_γ.

The extra model flexibility introduced by modelling tail behaviour and skewness is thus seen not to affect the propriety of the posterior distribution. As a consequence, the well-known result under Normal sampling holds in our much more general framework. Throughout the remainder of the article, we shall always assume $n \geq m + k$.

The following definition from Fernández and Steel (2000), concerning the design matrix X, is required to adequately characterise the existence of the marginal posterior moments of B.

Definition 1. Given an $n \times k$ full column-rank matrix X, the singularity index for column $l = 1,2,\ldots,k$ is defined as the largest number p_l ($0 \leq p_l \leq n-k$) such that there exists a $(k-1+p_l)\times k$ submatrix of X of rank $k-1$ that remains of rank $k-1$ after removing its lth column.

Clearly if X contains rows of zeros, then p_l is at least equal to the number of such rows for all $l = 1,2,\ldots,k$. Furthermore, $\max\{p_l,\ j = 1,2,\ldots,k\} = 0$ if and only if every $k\times k$ submatrix of X is non-singular. The singularity index p_l plays a fundamental role in the existence of posterior moments of B_{lj}.

Theorem 3. Consider the Bayesian model given in (11), (13) and (14) and $r > 0$. Let $\mathcal{N}^* = (\nu_0^*, \infty)$, $\nu_0^* \geq 0$. Then $E\left(|B_{lj}|^r|Y\right) < \infty$ if $r < \min\{n-m-k+1, m(n-k-p_l)+\nu_0^*\}$, with p_l the singularity index for column l of the design matrix X.

Theorem 3 only considers non-negative moments. Since the first negative moment of a Normally distributed random variable does not exist, the moments in Theorem 3 are always infinite for values of $r \leq -1$. This result is also a feature of inference under Normal sampling. The proof of Theorem 3 also shows that if $r \geq n-m-k+1$ there is no possibility for the moment to exist, regardless of the properties of the design matrix or the prior P_{ν^*}. Such a result is due to the uncertainty about B and Σ, rather than the remaining components of the model. However, both X and P_{ν^*} intervene in the sufficient condition stated in Theorem 3.

We now turn our attention to the posterior moments of $|\Sigma|$ of order $r/2 \geq 0$. For this quantity, the order up to which the posterior moments are finite does not depend on the design matrix or the distributions of P_γ and P_{ν^*} as is stated in the following theorem.

Theorem 4. Consider the Bayesian model given in (14), under the prior in (11) and (13), and $r \geq 0$. Then, $E\left(|\Sigma|^{\frac{r}{2}}|Y\right) < \infty$ if and only if $r < n-m-k+1$.

Note that if we impose a Dirac distribution on $\lambda_i = 1$, *i.e.* $p(\lambda_i = 1) = 1,\ i = 1, \ldots, n$, we obtain a regression model with skew-Normal disturbances. The results above apply to this model in the limit as $\nu_0^* \to \infty$. Also, if we set the components of γ equal to one, we obtain the symmetric versions of the distributions. We know from Theorem 1 that the results derived here also apply to the case of symmetric Student sampling. In that case, as explained in Subsection 2.2, the matrix O is no longer necessary for inference, and therefore we can set it to $O = I_m$, the $m \times m$ identity matrix.

3.3 Inference under skew-IStudent sampling

The regression framework introduced in the previous subsection implies a common tail behaviour for ϵ along all directions. Here we relax that assumption by allowing $f_{\nu_j}(\cdot),\ j = 1, \ldots, m$ to have different tail behaviour. In particular, we adopt the Bayesian regression model in (10)-(11) and (13) with

- For $j = 1, \ldots, m$, $f_{\nu_j}(\cdot)$ is the univariate Student distribution with ν_j df.
- P_{λ_i} is a Dirac distribution on $\lambda_i = 1,\ i = 1, \ldots, n$.
- $P_\nu = \prod_{j=1}^m P_{\nu_j}$.

The sampling distribution is then given by

$$p(y_i|B, O, U, \nu, \gamma) = |U|^{-1} \prod_{j=1}^{m} \frac{2}{\gamma_j + \frac{1}{\gamma_j}} f_{\nu_j}\left(d_{ij}\ \gamma_j^{-\mathrm{sign}(d_{ij})}\right), \tag{15}$$

where $f_{\nu_j}(\cdot)$ is given in (8) and d_{ij} is as in (14). This defines the m-dimensional skew-IStudent with location $B'x_i$, transformation OU, skewness parameter γ and df vector $\nu = (\nu_1, \ldots, \nu_m)'$.

For the propriety of the posterior distribution, we obtain the following.

Theorem 5. Consider n independent replications from the sampling model in (15) under the prior in (11) and (13). If for any $j = 1, \ldots, m$, $P(\nu_j \leq m-1) = 0$ and $n \geq m + k$, then the posterior distribution is proper for any P_γ.

The requirement that $P(\nu_j \leq m-1) = 0$ can be restrictive, especially if the dimension of the problem is large. However, for reasonably small m, the restriction is unlikely to cause much harm, as only distributions with extremely heavy tails are excluded. In what follows, we always assume that P_ν complies with the sufficient condition in Theorem 5.

The following theorems focus on the existence of the moments of B and $|\Sigma|$.

Theorem 6. Consider the Bayesian model given in (11), (13) and (15) and $r > 0$. Let $\mathcal{N} = (\nu_0, \infty)$ be the common support of P_{ν_j}, $j = 1, \ldots, m$. Then we obtain that $E\left(|B_{lj}|^r|Y\right) < \infty$ if $r < \min\{n-m-k+1, m(n-k-p_l-1)+\nu_0+1\}$, with p_l the singularity index for column l of the design matrix X.

Theorem 7. Consider the Bayesian model given in (15), under the prior in (11) and (13), and $r \geq 0$. Then, $E\left(|\Sigma|^{\frac{r}{2}}|Y\right) < \infty$ if $r < n-m-k+1$.

If we consider the special case where the components of γ are set to one, we obtain sampling under the Independent Student (IStudent) distribution. However, as the product of univariate Student distributions is not in the spherical class, it is still necessary to consider the orthogonal matrix O. To our knowledge, this sampling model has not been analysed in the literature, even under symmetry. Thus, this subsection also introduces a novel class of symmetric heavy-tailed distributions and analyses its properties in a Bayesian regression context.

3.4 Completing the prior specification

Having already specified the prior structure and the prior distributions for B and O and U, the Bayesian models become fully specified by assigning proper prior distributions for γ, ν^* and ν.

We assume that the components of $\gamma \in \Re_+^m$ are independently distributed according to a common logNormal distribution, *i.e.* $\log(\gamma_j) \sim N(0, s^2)$, $j = 1, \ldots, m$. This centers the prior over symmetry and implies that for any two constants such that $1 < \gamma_a < \gamma_b$, we have $P[\gamma_j \in (\gamma_a, \gamma_b)] = P\left[\gamma_j \in \left(\frac{1}{\gamma_b}, \frac{1}{\gamma_a}\right)\right]$, thus treating positive and negative skewness symmetrically in the prior.

For ν^* and the components of ν we use an Exponential prior with parameter $d > 0$, restricted to $\mathcal{N}^* = \mathcal{N} = (\max\{3, m-1\}, \infty)$, allowing at the same time the use of improper priors and calculation of the third moments, necessary to calculate the Mardia measure of skewness. In addition, it will, in most practical situations, avoid the problems of posterior nonexistence with point observations pointed out in Fernández and Steel (1999) for symmetric multivariate Student regression models.

Finally, it is necessary to choose values for the two hyperparameters defined above: s and d. We set the first of these to one, corresponding to a rather vague prior on γ, and the second to 0.1 as in Fernández and Steel (1998).

4 Numerical Implementation

Inference with the Bayesian models introduced in Section 3 requires numerical methods. Here we conduct inference using Markov chain Monte Carlo (MCMC) methods, in particular hybrid samplers with both Metropolis-Hastings and Gibbs components.

As the univariate Student-t distribution can be expressed as a mixture of Normals, (15) can be written as

$$p(y_i|B,O,U,\nu,\gamma) = |U|^{-1} \prod_{j=1}^{m} \frac{2}{\gamma_j + \frac{1}{\gamma_j}} \int_{\Re_+} \lambda_{ij}^{\frac{1}{2}} \, \phi\left(\lambda_{ij}^{\frac{1}{2}} \, d_{ij} \, \gamma_j^{-\mathrm{sign}(d_{ij})}\right) p_G\left(\lambda_{ij} \middle| \frac{\nu_j}{2}, \frac{\nu_j}{2}\right) d\lambda_{ij}. \tag{16}$$

This illustrates a fundamental difference between the skew-Student and the skew-IStudent sampling models. In the skew-Student in (14), each observation y_i has its corresponding mixing parameter λ_i, $i = 1, \ldots, n$, pairwise independent and identically distributed given ν^*. For the skew-IStudent model, each observation y_i has its vector of independent mixing parameters $\lambda_i = (\lambda_{i1}, \ldots, \lambda_{im})'$, with different distributions for each element. Thus, even if $\nu_j = \nu^*, j = 1 \ldots, m$, the two models are still quite different. For the skew-IStudent model, we conduct inference on $(B, O, U, \gamma, \lambda, \nu \mid Y)$, where $\lambda = (\lambda_1, \ldots, \lambda_n)'$ while for skew-Student sampling, we merely replace ν by ν^*.

Most steps in the sampler are fairly standard, with one exception: the step to draw O, which will be explained below. For both models, the components of λ are independent given the other parameters, and can directly be sampled from Gamma distributions. Drawings from the conditional posterior distributions for ν and ν^* are generated using a rejection sampler. We use individual random-walk Metropolis-Hastings samplers for B, O, U and γ, common to both models. For the components of B, the off-diagonal elements of U and the logarithm of the components of γ we use a Normal proposal, while we use a half-Normal proposal distribution for the diagonal elements of U. Throughout, we update one component at a time.

4.1 Sampling O

Sampling orthogonal matrices $O \in \mathcal{O}^m$ directly is extremely complicated. Thus we use a reparameterisation of O, more suitable for sampling.

Let us first define $(\theta^2, \ldots, \theta^m)' \in \Theta^2 \times \ldots \Theta^m$, where $\theta^j = (\theta^j_1, \ldots, \theta^j_{j-1})$ and Θ^j is as defined in Appendix B.4.

Appendix B shows that if the $v^j = (v^j_1, \ldots v^j_j)'$ are such that,

- $v^j_1 = \sin(\theta^j_1)$
- $v^j_i = \prod_{l=1}^{i-1} \cos(\theta^j_l) \times \sin(\theta^j_i), \quad i < j$
- $v^j_j = \prod_{l=1}^{j-1} \cos(\theta^j_l)$,

the matrix $H_{\theta^j} = I_j - 2v^j \left(v^j\right)'$ and $O^m_{\theta^j}$ is defined as

$$O^m_{\theta^j} = \begin{pmatrix} I_{m-j} & 0 \\ 0 & H_{\theta^j} \end{pmatrix},$$

then we can express any $m \times m$ orthogonal matrix $O \in \mathcal{O}^m$ as $O = O^m_{\theta^m} \times \ldots \times O^m_{\theta^2}$.

We can then sample O easily by sampling in turn each component of each θ^j, $j = 2, \ldots, m$. For each one of those, we sample from a Normal random walk proposal distribution, restricted so that $\theta^j \in \Theta^j$ and, thus, the proposed orthogonal matrix is in $\mathcal{O}^m$.

5 Examples

Along with the most general models, incorporating both skewness and fat tails - skew-Student and skew-IStudent, we also consider simpler alternatives: Student, IStudent, skew-Normal and Normal, which are nested in at least one of the more general models. The Student, IStudent and the Normal models assume symmetry ($\gamma = 1$), while the latter model and the skew-Normal do not allow for heavy tails. The prior distributions for the parameters of these models are compatible with those in the more general ones.

Here, we do not present any comparison with regression models based on other classes of skewed distributions. Up to our knowledge, no other method has been shown to allow for inference under an improper prior structure compatible with (11) and (13). We refer the interested reader to Ferreira and Steel (to appear), where a comparison between our methodology and the one in Sahu et al. (2003) is presented under a proper prior.

In both applications, inference was conducted using every tenth of 100,000 realisations from the Markov chain described in Section 4, after discarding the first 20,000 samples (a burn-in sufficient for convergence in all cases).

In what follows, we present posterior and predictive inference. Model comparison is also provided through Bayes factors. Estimates of marginal likelihoods are obtained using the p_4 measure in Newton and Raftery (1994), with their δ set to 0.1.

5.1 Firm size data

The study of the distribution of firm size is an important problem in economics, generating a substantial research effort. The most widely studied model is the one initially developed in Gibrat (1931), known as Gibrat's law or the law of proportionate effect, where a Lognormal distribution is assumed. The history, implications and developments of Gibrat's law are reviewed in Sutton (1997).

Here we analyse the size distribution of small and medium businesses, registered in the United Kingdom, using data from the Cambridge Centre for Business Research SME Dataset, 1987-1995 (Cosh et al. 2002). The data pertain to $n = 539$ exporting firms for which the volume of exports (in thousands of pounds) and the number of employees was registered in 1987. To study the validity of Gibrat's law on these data, we define the size variables Exports and Employment as the natural logarithm of the original data. We use a location-scale model (*i.e.* with only a constant term).

Figure 5 (a) presents the marginal posterior densities of $\gamma = (\gamma_1, \gamma_2)'$ for the skew-Normal model, shown to be the most adequate of all models, together with the prior distribution for $\gamma_j,\ j = 1, 2$. Both components of γ are seen to be different from one, implying skewness. We stress that from the analysis of Figure 5 (a) it is not possible to infer that one of the variables of interest is positively skewed, while the other is negatively skewed. As shown in Subsection 2.2, the shape of the distribution is determined jointly by γ and $A = OU$. In Figure 5 (b) we show a grayscale plot of the posterior pdf of the basic axes, as defined in Subsection 2.2.3, for the skew-Normal model. In the plot, darker tones correspond to higher posterior densities. Figure 5 (b) illustrates the need of considering the direction of skewness. In this particular example, the direction of the basic axes is rather different from the coordinate axes defined by the variables.

For all models allowing heavier tails, the posterior distributions of the df are concentrated on high values, corresponding to tail behaviour close to the one of the Normal distribution.

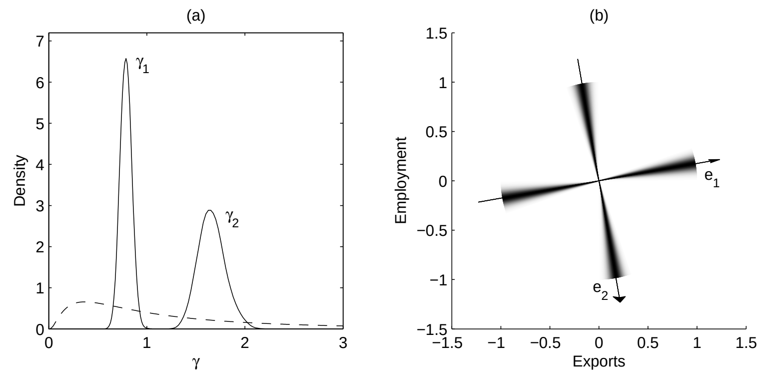

Figure 5: (a) Estimates of the marginal pdfs of the components of $\gamma = (\gamma_1, \gamma_2)$ for the skew-Normal (solid line) model together with the prior pdf for the elements of γ (dashed line); (b) grayscale plot of the posterior pdf of the basic axes, as defined in Section 2, for the skew-Normal model.

The shape of the distribution is illustrated in Figure 6, which presents a contour plot of the predictive distribution from the skew-Normal model, superimposed over the observations indicated by the dots. It is clear that the distribution of the data is not symmetric, but the predictive contours seem to capture the main characteristics quite well.

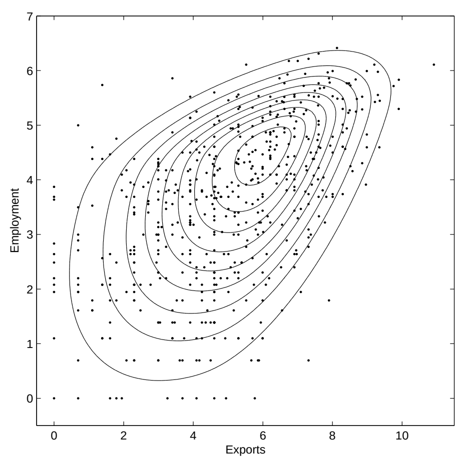

Figure 6: Contour plot of the predictive distribution of the variables estimated with the skew-Normal model, overlapping the data observations, denoted by the dots.

A formal comparison of the six models using Bayes factors is given in Table 1. A value larger than unity in entry (i, j) indicates support in favour of model i. The analysis reveals that the models allowing for skewness are far superior to the symmetrical ones (all Bayes factors exceed 1000). Also, tail behaviour of the Normal models seems more adequate, although evidence here is less overwhelming. It is interesting to note that the (skew)-Student beats the (skew)-IStudent which is consistent with the parsimony preference of Bayes factors. In summary, the skew-Normal alternative is the most suitable for the distribution of size of the businesses we study.

Our results suggest the inadequacy of Gibrat's law for these data. Whereas tail behaviour is adequately described by the law, the asymmetry of the distribution is clearly not. A more complete analysis of the skewness of distributions of firm size is presented in Ferreira and Steel (to appear).

Table 1: Bayes factors for firm size data. Entries indicate support in favour of the model in the row versus that in the column

	skew-Student	Student	skew-Normal	Normal	skew-IStudent	IStudent
skew-Student	1	> 1000	0.02	> 1000	2.2	> 1000
Student		1	< 0.001	0.11	< 0.001	4.4
skew-Normal			1	> 1000	90	> 1000
Normal				1	< 0.001	40
skew-IStudent					1	> 1000

5.2 Australian Institute of Sport Data

For our second example we use a dataset from the Australian Institute of Sport. In particular, we study the distribution of four biomedical variables: body mass index (BMI), sum of skin folds (SSF), percentage of body fat (PBF), and lean body mass (LBM). The data were collected for 202 athletes at the Australian Institute of Sport and are described in Cook and Weisberg (1994). Besides a constant term we use information on three covariates: red cell count (RCC), white cell count (WCC) and plasma ferritin concentration (PFC). In order to compare the influence of the covariates, the data was normalised to have mean zero and variance one. These data, without the covariates, have been used previously in the context of skewed distributions in, *i.a.*, Azzalini and Capitanio (1999, 2003).

Figure 7 (a) presents the marginal posterior pdfs of the elements of γ for the model that proved to be most adequate - the skew-IStudent model, together with the prior pdf. For all but one components of γ, the pdfs have low mass near unity, implying that the data requires that at least three components in the linear transformation are skewed. Figure 7 (b) exhibits the posterior pdf of Mardia's measure of skewness for the skew-IStudent and also the prior distribution for that quantity. The posterior pdf of $\beta_{1,4}$ has most of its mass concentrated away from zero, implying that the distribution of the quantities of interest is asymmetric. We note that by assuming our fairly uninformative prior on γ, the prior on $\beta_{1,4}$ puts substantial mass on asymmetric distributions, but also retains mass on low values corresponding to symmetric distributions. The posterior distributions of γ and $\beta_{1,4}$ are fairly robust to changes of the prior.

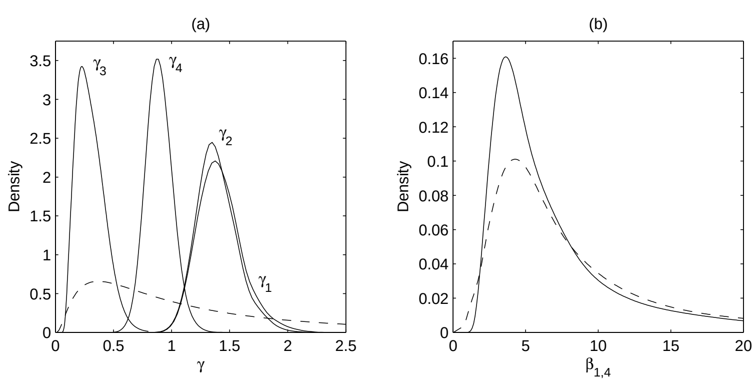

Figure 7: (a) marginal posterior pdfs for the components of γ for the skew-IStudent model (solid) together with the prior pdf (dashed); (b) posterior pdf of Mardia's measure of skewness for the skew-Student (solid) and the prior pdf for the same quantity (dashed)

We now assess the relevance of the covariates on the values of the variables of interest. Figures 8 (a)-(d) present the posterior distributions of the coefficients of B for the intercept, RCC, WCC and PFC, respectively. In most cases, the covariates are shown to have an effect on the distribution of the variables, particularly for RCC. BMI does not seem to need any of the covariates, but all regressors intervene crucially in modelling SSF. The posterior distribution of the regression coefficients is quite different from the prior distribution, which is improper uniform.

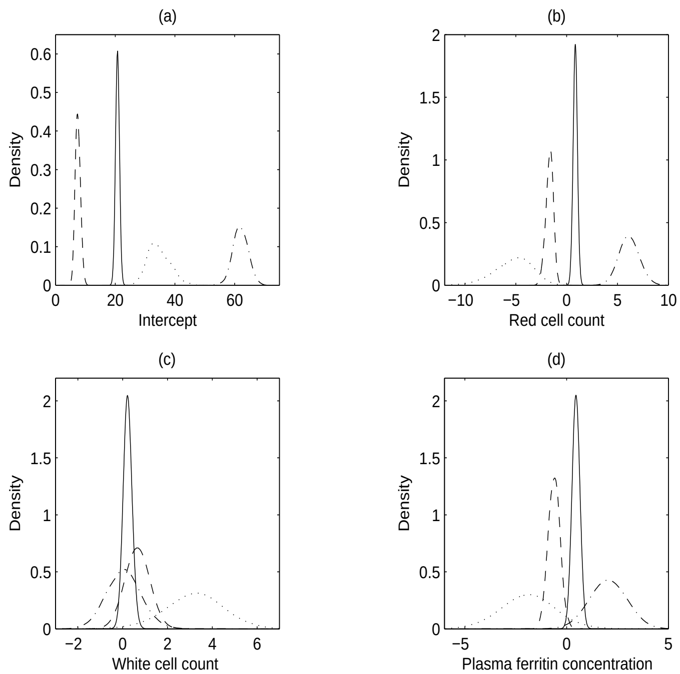

Figure 8: (a)-(d) posterior distributions of the coefficients of B for the intercept, RCC, WCC and PFC, respectively, corresponding to BMI (solid), SSF (dotted), BFAT (dashed) and LBM (dot-dashed), evaluated for the skew-Student model

Unlike in the firm size example, the distribution of the biomedical measurements has heavier than Normal tails. Figure 9 presents the posterior density for the df for the skew-IStudent model, together with the prior distribution. Some components require much heavier tails than others with the medians of ν_j given by 10.7, 16.2, 5.7 and 13.4, $j = 1, \ldots, 4$. The skew-Student model leads to a median value of ν^* equal to 15.8. Both models lead to heavier tails when we impose symmetry. Thus, if we (wrongly) impose symmetry, the skewness in the data is partly misinterpreted as fat tails.

Table 2 compares the models using Bayes factors. We conclude that the skew-IStudent model is the most favoured model for these data, with the skew-Student model a distant second. As in the previous example, a large difference exists between the adequacy of the skewed models and the others (Bayes factors larger than 1000). There is also strong evidence in favour of heavy tails, but interestingly, the IStudent tails receive a lot more support from the data than the Student tails. This is partly due to the differences between the ν_js in Figure 9, but, as explained in Section 4, other differences exist between these models. In summary, both skewness and heavy tails are strongly supported, which is in line with the findings of Azzalini and Capitanio (2003) in the context of a location-scale model.

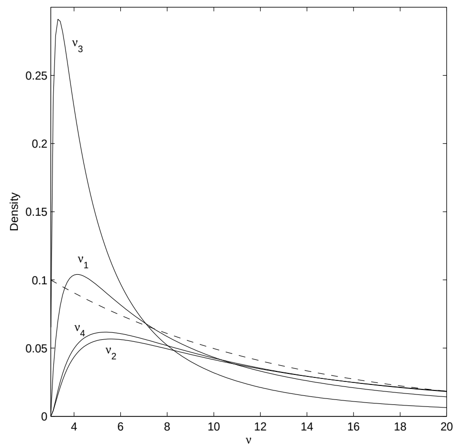

Figure 9: Posterior density for $\nu_1, \ldots, \nu_4$ for the skew-IStudent model (solid line), together with the prior pdf (dashed). Note the truncation at three.

Table 2: Bayes factors for Australian Institute of Sport data. Entries indicate support in favour of the model in the row versus that in the column

	skew-Student	Student	skew-Normal	Normal	skew-IStudent	IStudent
skew-Student	1	> 1000	2.6	> 1000	< 0.001	> 1000
Student		1	< 0.001	26	< 0.001	< 0.001
skew-Normal			1	> 1000	< 0.001	> 1000
Normal				1	< 0.001	< 0.001
skew-IStudent					1	> 1000

6 Conclusion

In this article we present a novel general method for the construction of skewed multivariate distributions. Based on linear transformations of univariate skewed distributions, the method we introduce is quite flexible, *i.e.* it imposes few restrictions on the form of the distribution.

In particular, we use linear transformations of independent and univariate random quantities with skewed distributions as in Fernández and Steel (1998). The generated class of distributions has a number of appealing characteristics. Moment calculation is always straightforward if the moments of the underlying univariate distribution are available, and mean, variance and skewness can be modelled independently. Also, unlike other classes of skewed distributions proposed in the literature, our method makes no use of conditioning arguments, which require the calculation of cumulative distributions functions. This aspect can be quite relevant, especially for certain distributions and for high dimensions. A drawback of our proposal is that the class of skewed distributions is in general not closed under linear transformations or marginalisation.

We provide results on inference with these skewed distributions in a Bayesian regression model, under commonly used improper priors, and show that the extra flexibility induced by skewness does not have any impact on the existence of the posterior distribution, or even on the existence of posterior moments of the parameters. Further, we introduce two classes of skewed and heavy-tailed multivariate regression models, skew-IStudent and skew-Student, and establish results on posterior propriety and existence of posterior moments. One of these classes (the skew-IStudent) is novel even under symmetry.

This article uses as the underlying univariate skewed distribution the class introduced in Fernández and Steel (1998). However, the proposed method to generate multivariate distributions is more generally applicable and is not restricted to this class.

We introduce MCMC samplers to conduct inference with these skewed and heavy-tailed linear regression models and examine two applications: firm size and biomedical measurements. We compare skewed models with symmetric models and heavy tails with Normal ones. In both applications we find strong evidence in favour of skewed distributions.

Inference under the models we study in this article is quite feasible even with a practically relevant number of dimensions and/or predictors. As an illustration, our MCMC samplers applied to the four-dimensional Australian Institute of Sport data, with four regressors, implemented using Matlab on a Pentium 4, 2.7 GHz PC, run at around 230 and 190 thousand iterations per hour for the skew-Student and skew-IStudent models, respectively. The Matlab implementation is freely available from http://www.warwick.ac.uk/go/jtferreira/mskew.html.

A Proofs and Auxiliary Results

Outlines of proofs are provided below, with extensive referring to the literature.

Proof of Lemma 1. See Graybill (1983, p. 210-211). □

Definition 2. For $\lambda \in \Re^n_+$, define $\lambda_{(1)}, \ldots, \lambda_{(n)}$ to be the ordered λ_i's.

Definition 3. For $\lambda \in \Re^n_+$, define $\{\lambda_{m_1}, \ldots \lambda_{m_k}\}$ as the set of λ_i's that satisfies $\prod_{i=1}^k \lambda_{m_i} = \max\{\prod_{i=1}^k \lambda_{s_i} : 1 \leq s_1 \leq \ldots \leq s_n \leq n$ and $|[x_{s_1}, \ldots, x_{s_n}]| \neq 0\}$, where $[x_{s_1}, \ldots x_{s_n}]$ denotes the submatrix of X corresponding to the observations $y_{s_1}, \ldots y_{s_n}$.

Lemma 2. Let $\Lambda = \text{diag}(\lambda)$, the Euclidean norm of $B(\lambda) = (X'\Lambda X)^{-1}X'\Lambda Y$ is bounded above by a finite constant $C(X, Y)$.

Proof. The proof follows the one of Lemma 2 in Fernández and Steel (2000) taking one column of $B(\lambda)$ at a time as $b(\lambda)$ in the notation of that proof. □

Lemma 3. Let $S(\lambda) = Y'\{\Lambda - \Lambda X(X'\Lambda X)^{-1}X'\Lambda\}Y$, then for all $Y \in \Re^{n\times m}$ barring a set of Lebesgue measure zero, the expression $|S(\lambda)|$ has upper and lower bounds proportional to λ_b^m where $\lambda_b = \max\{\lambda_i : i \neq m_1, \ldots, m_k\}$.

Proof. Let $L = (X \ : \ Y)$. After some matrix manipulation we have,

$$|L'\Lambda L| = \left|\begin{pmatrix} X'\Lambda X & X'\Lambda Y \\ Y'\Lambda X & Y'\Lambda Y \end{pmatrix}\right| = |X'\Lambda X||Y'\Lambda Y - Y'\Lambda X(X'\Lambda X)^{-1}X'\Lambda Y|,$$

and, therefore, $|S(\lambda)| = |L'\Lambda L|\{|X'\Lambda X|\}^{-1}$. Application of the Binet-Cauchy formula (Gantmacher

(1959, p.9) leads to

$$|S(\lambda)| = \{|X'\Lambda X|\}^{-1} \sum_{1 \le s_1 \le \ldots \le s_{k+m}} \left(\prod_{i=1}^{k+m} \lambda_{s_i} \right) \left| \begin{pmatrix} x_{s_1} \ldots x_{s_{k+m}} \\ y_{s_1} \ldots y_{s_{k+m}} \end{pmatrix} \right|^2 .$$

In combination with the proof of Lemma 1 in Fernández and Steel (2000) we see that $|S(\lambda)|$ has upper and lower bounds proportional to $\prod_{i=k+1}^{k+m} \lambda_{s_i}$ and this last quantity is not larger than λ_b^m. □

Lemma 4. For all $Y \in \Re^{n \times m}$ barring a set of Lebesgue measure zero and for all $j = 1, \ldots, m$ the jth diagonal element of $S(\lambda)$ has upper and lower bounds proportional to λ_b.

Proof. The proof is as for Lemma 3 with $L = (X : Y_{\cdot j})$, where $Y_{\cdot j}$ is the jth column of Y. □

Lemma 5. The lth diagonal element of $(X'\Lambda X)^{-1}$ has upper and lower bounds proportional to $1/\lambda_{a^l}$, where $\lambda_{a^l} = \min\{\lambda_a : a \in \{m_1, \ldots, m_k\}$ and $|({}^l x_{m_i} : m_i \neq a)| \neq 0\}$, ${}^l x_{m_i}$ denotes the vector x_i without its lth element and $({}^l x_{m_i} : m_i \neq a)$ is the $(k-1) \times (k-1)$ matrix obtained from $({}^l x_{m_i}, \ldots, {}^l x_{m_k})$ after removing ${}^l x_a$.

Proof. See the proof of Lemma 4 in Fernández and Steel (2000). □

Lemma 6. Let $p_S(\epsilon|\nu^*)$ denote the multivariate Student pdf with ν^* df and let $p_{IS}(\epsilon|\nu = (\nu_1, \ldots, \nu_m)')$ denote the multivariate pdf of the independent Student components distribution with df vector ν. Thus,

$$p_S(\epsilon|\nu^*) = \frac{\Gamma\left(\frac{\nu^*+m}{2}\right)}{\Gamma\left(\frac{\nu^*}{2}\right)(\pi\nu^*)^{\frac{m}{2}}} \left[1 + \frac{\epsilon'\epsilon}{\nu^*}\right]^{\frac{\nu^*+m}{2}}$$

and

$$p_{IS}(\epsilon|\nu) = \prod_{j=1}^{m} f_{\nu_j}(\epsilon_j), \tag{17}$$

where $f_\nu(\epsilon_j)$ is the univariate Student pdf with ν df as in (8). Then, there exists a positive constant K such that, for any $\epsilon \in \Re^m$, we have that $p_{IS}(\epsilon|\nu) \le K p_S(\epsilon|\nu^\bullet - m + 1)$, where $\nu^\bullet = \min\{\nu_1, \ldots, \nu_m\}$.

Proof. If $\nu^a \le \nu^b$ it is clear that there exist $G_1(\nu^a, \nu^b) > 0$ such that $f_{\nu^b}(\epsilon) \le G_1(\nu^a, \nu^b) f_{\nu^a}(\epsilon)$. We define

$$G_1(\nu^a, \nu^b) = \max_\epsilon \frac{f_{\nu^b}(\epsilon)}{f_{\nu^a}(\epsilon)} = \frac{(\nu^b)^{\nu_b/2}}{(\nu^a)^{\nu_a/2}} \frac{\Gamma\left(\frac{\nu^b+1}{2}\right)\Gamma\left(\frac{\nu^a}{2}\right)}{\Gamma\left(\frac{\nu^a+1}{2}\right)\Gamma\left(\frac{\nu^b}{2}\right)} \frac{(\nu^a+1)^{\frac{\nu^a+1}{2}}}{(\nu^b+1)^{\frac{\nu^b+1}{2}}}, \tag{18}$$

which is finite if $\nu^a, \nu^b > 0$. Using this result, we can find a finite $K_1(\nu^\bullet) > 0$, such that $p_{IS}(\epsilon|\nu) \le K_1(\nu^\bullet) p_{IS}(\epsilon|\nu_m^\bullet)$, where $\nu_m^\bullet$ denotes the m-dimensional vector where all elements equal $\nu^\bullet$.

Now we prove that there exists $K_2(\nu^\bullet) > 0$ such that $p_{IS}(\epsilon|\nu_m^\bullet) \le K_2(\nu^\bullet) p_S(\epsilon|\nu^\bullet - m + 1)$. If $\|\epsilon\| = \rho$ is fixed, then $p_{IS}(\epsilon|\nu_m^\bullet)$ has its maximum value at those ϵ for which only one component differs from zero. For simplicity assume that $\epsilon = (\rho, 0, \ldots, 0)'$. Then we can state

$$G_2(\nu^*, \nu^\bullet) = \max_\rho \frac{p_{IS}(\epsilon|\nu_m^\bullet)}{p_S(\epsilon|\nu^*)} = \left(\frac{\nu^*}{\nu^\bullet}\right)^{\frac{m}{2}} \left(\frac{\Gamma\left(\frac{\nu^\bullet}{2}\right)}{\Gamma\left(\frac{\nu^\bullet+1}{2}\right)}\right)^m \frac{\Gamma\left(\frac{\nu^*+m}{2}\right)}{\Gamma\left(\frac{\nu^*}{2}\right)} \frac{\left(1 + \frac{\rho^2}{\nu^\bullet}\right)^{\frac{\nu^\bullet+1}{2}}}{\left(1 + \frac{\rho^2}{\nu^*}\right)^{\frac{\nu^*+m}{2}}},$$

which exists, *i.e.* is finite, if and only if $\nu^* \leq \nu^\bullet - m + 1$ and its limit when $\nu^\bullet$ tends to infinity always exists. Thus, $K = \max_{\nu \in \mathcal{N}} K_1(\nu) K_2(\nu)$, which is necessarily finite and bounded. $\square$

Lemma 7. Let $\nu = (\nu_1, \ldots, \nu_m)'$, and let the distribution of ν, P_ν, be proper on $(m-1, \infty)^m$. Then, the distribution of $\nu^\bullet = \min\{\nu_1, \ldots, \nu_m\}$ is also proper on $(m-1, \infty)$.

Proof. Trivial. $\square$

Proof of Theorem 1

Under the sampling distribution (10) and the prior in (11) we have that

$$E\left(|\Sigma|^{\frac{r}{2}} \prod_{l=1}^{p} |B_l|^{r_l} \,\middle|\, Y\right) = \int_{\Re^{k\times m}\times\mathcal{O}^m\times\mathcal{U}^m\times\mathcal{L}^n\times\Re_+^m\times\mathcal{N}^m} |U|^{r-n} \prod_{l=1}^{p} |B_l|^{r_l} \times$$

$$\times \prod_{i=1}^{n} \lambda_i^{\frac{m}{2}} p\left\{\lambda_i^{\frac{1}{2}} [(OU)']^{-1}[y_i - g_i(B)] \middle| \gamma, f_\nu\right\} dP_B dP_O dP_U dP_\lambda dP_\gamma dP_\nu. \quad (19)$$

Using the reasoning at the end of Section 2.2, writing $A = OU$ and expanding the domain of O from $\mathcal{O}^m$ to O^m, the set of $m \times m$ orthogonal matrices, (19) is proportional to

$$\int_{\Re^{k\times m}\times\mathcal{NS}^m\times\mathcal{L}^n\times\Re_+^m\times\mathcal{N}^m} \|A\|^{r-n} \prod_{l=1}^{p} |B_l|^{r_l} \prod_{i=1}^{n} \lambda_i^{\frac{m}{2}} p\left\{\lambda_i^{\frac{1}{2}} (A')^{-1}[y_i - g_i(B)] | \gamma, f_\nu\right\} dP_B dP_A dP_\lambda dP_\gamma dP_\nu, \quad (20)$$

where $\mathcal{NS}^m$ denotes the set of non-singular $m \times m$ matrices. The equation above is bounded by

$$2^{nm} \int_{\Re^{k\times m}\times\mathcal{NS}^m\times\mathcal{L}^n\times\Re_+^m\times\mathcal{N}^m} \|A\|^{r-n} \prod_{l=1}^{p} |B_l|^{r_l} \prod_{j=1}^{m} \left(\frac{1}{\gamma_j + \frac{1}{\gamma_j}}\right)^n \times$$

$$\times \prod_{i=1}^{n} \lambda_i^{\frac{m}{2}} p\left\{\lambda_i^{\frac{1}{2}} \Gamma_h^{-1} (A')^{-1}[y_i - g_i(B)] | 1_m, f_\nu\right\} dP_B dP_A dP_\lambda dP_\gamma dP_\nu,$$

where 1_m denotes the m-dimensional vector of ones, and with Γ_h the diagonal matrix with entries $h(\gamma_j)$, $j = 1, \ldots m$, where

$$h(\gamma_j) = \begin{cases} \min\{\gamma_j, \gamma_j^{-1}\} & \text{for the lower bound} \\ \max\{\gamma_j, \gamma_j^{-1}\} & \text{for the upper bound.} \end{cases}$$

Applying Fubini's theorem, we consider first the integral with respect to A. Transforming A to $H = \Gamma_h A$ leads to upper and lower bounds for (20) proportional to the product of

$$\int_{\Re_+^m} |\Gamma_h|^{n-r} \prod_{j=1}^{m} \left(\frac{1}{\gamma_j + \frac{1}{\gamma_j}}\right)^n dP_\gamma \quad (21)$$

and

$$\int_{\Re^{k\times m}\times\mathcal{NS}^m\times\mathcal{L}^n\times\mathcal{N}^m} \|H\|^{r-n} \prod_{l=1}^{p} |B_l|^{r_l} \prod_{i=1}^{n} \lambda_i^{\frac{m}{2}} p\left\{\lambda_i^{\frac{1}{2}} (H')^{-1}[y_i - g_i(B)] | 1_m, f_\nu\right\} dP_B dP_H dP_\lambda dP_\nu. \quad (22)$$

Clearly, for $h(\gamma_j) = \min\{\gamma_j, \gamma_j^{-1}\}$, the integral in (21) is strictly positive. If $h(\gamma_j) = \max\{\gamma_j, \gamma_j^{-1}\}$ than that integral is smaller than 1 as $h(\gamma_j) > 1$ and $r \geq 0$. Making use of Lemma 1, writing $H = OU$

and again using Fubini's theorem, we have that (22) is proportional to

$$\int_{\Re^{k\times m}\times\mathcal{O}^m\times\mathcal{U}^m\times\mathcal{L}^n\times\mathcal{N}} |U|^{r-n}\prod_{l=1}^{p}|B_l|^{r_l}\prod_{i=1}^{n}\lambda_i^{\frac{m}{2}}p\left\{\lambda_i^{\frac{1}{2}}O(U')^{-1}[y_i-g_i(B)]|1_m,f_\nu\right\}dP_BdP_OdP_UdP_\lambda dP_\nu, \tag{23}$$

which coincides with (19) when the errors are symmetrically distributed. □

Proof of Theorem 2

Follows from either of the proofs of Theorems 3 and 4. □

Proof of Theorem 3

As a result of Theorem 1, the existence of posterior moments for the Bayesian model in (11) and (14) is equivalent to the existence of moments under symmetrically distributed errors. Also, as the Normal distribution is closed under linear transformations, proving Theorem 3 is equivalent to proving that

$$\int_{\Re^{k\times m}\times\mathcal{C}^m\times\Re_+^n\times\mathcal{N}^*}|B_{lj}|^r|\Sigma|^{-\frac{n+m+1}{2}}|\Lambda|^{\frac{m}{2}}\exp\left\{-\frac{1}{2}\mathrm{tr}\left[\Sigma^{-1}(Y-XB)'\Lambda(Y-XB)\right]\right\}dBd\Sigma dP_{\lambda,\nu^*} \tag{24}$$

is finite for $r<\min\{n-m-k+1,m(n-k-p_l)+\nu_0^*\}$.

After some algebraic manipulation the existence of $E(|B_{lj}|^r|Y)$ is equivalent to

$$\int_{\Re\times\Re_+\times\Re_+^n\times\mathcal{N}^*}|B_{lj}|^r f_N\left[B_{lj}|B(\lambda)_{lj},\Sigma_{jj}\left(X'\Lambda X\right)_{ll}^{-1}\right]f_G\left[\Sigma_{jj}^{-1}|\frac{n-k-m+1}{2},\frac{S(\lambda)_{jj}}{2}\right]\times$$

$$\times|\Lambda|^{\frac{m}{2}}|X'\Lambda X|^{-\frac{m}{2}}|S(\Lambda)|^{-\frac{n-k}{2}}d\Sigma_{jj}^{-1}dP_{\lambda,\nu^*}<\infty, \tag{25}$$

with $B(\lambda)$ as defined in Lemma 2 and where $f_N(\cdot|\mu,\sigma^2)$ denotes the Normal pdf with mean μ and variance σ^2 and $f_G(\cdot|a,b)$ is the pdf of a Gamma distribution with shape a and precision b.

First we solve the integral with respect to B_{lj}, denoted by I_B. Making the variable transformation from B_{lj} to $b=B_{lj}-B(\lambda)_{lj}$, we have that

$$I_B=\int_{\Re}|b-B(\lambda)_{lj}|^r f_N\left\{b|0,\Sigma_{jj}\left[\left(X'\Lambda X\right)^{-1}\right]_{ll}\right\}db.$$

We now find a lower and an upper bound for I_B which lead to bounds of the integral (25). That if $r\geq n-m-k+1$, the integral in (25) is unbounded can be shown using a similar reasoning as in the proof of Theorem 2 (i) in Fernández and Steel (2000).

Now, from Lemma 2 we have that, for some positive quantity $C(X,Y)$, $|B(\lambda)_{lj}|\leq C(X,Y)$, $l=1,\ldots,k,\ j=1,\ldots,m$. Therefore it follows that

$$|b-B(\lambda)_{lj}|^r\leq 2^r|C(X,Y)|^r+2^r|b|^r.$$

As a consequence, in order to prove Theorem 3 we need to show the existence of

$$\int_{\Re\times\Re_+\times\Re_+^n\times\mathcal{N}^*}|b|^q f_N\left[b|b,\Sigma_{jj}\left(X'\Lambda X\right)_{ll}^{-1}\right]f_G\left[\Sigma_{jj}^{-1}|\frac{n-k-m+1}{2},\frac{S(\lambda)_{jj}}{2}\right]\times$$

$$\times|\Lambda|^{\frac{m}{2}}|X'\Lambda X|^{-\frac{m}{2}}|S(\Lambda)|^{-\frac{n-k}{2}}dbd\Sigma_{jj}^{-1}dP_{\lambda,\nu^*}. \tag{26}$$

for $q = 0$ and $q = r$. Also, note that for $q = 0$ this proof covers the proof of Theorem 2.

After integrating out b and Σ_{jj}, we are left with the integral

$$\int_{\Re_+^n \times \mathcal{N}^*} \left\{(X'\Lambda X)_{ll}^{-1}\right\}^{\frac{q}{2}} |\Lambda|^{\frac{m}{2}} |X'\Lambda X|^{-\frac{m}{2}} |S(\lambda)|^{-\frac{n-k}{2}} \left\{S(\lambda)_{jj}\right\}^{\frac{q}{2}} dP_{\lambda,\nu^*}. \tag{27}$$

We decompose the domain of integration $\Re_+^n$ into $n!$ possible orderings of $\{\lambda_1, \ldots, \lambda_n\}$. In each of these regions we identify $\lambda_{m_1}, \ldots, \lambda_{m_k}$ (Definition 3), λ_b (Lemma 3) and λ_{a^l} (Lemma 5). Given one of these ordering and applying the previous lemmas as well as Lemma 1 in Fernández and Steel (2000), we obtain an upper and lower bound of the integrand in (27) proportional to

$$F(\lambda) = \frac{\prod_{s \neq m_1, \ldots, m_k} \lambda_s^{\frac{m}{2}}}{\lambda_{a^l}^{\frac{q}{2}} \lambda_b^{\frac{m(n-k)-q}{2}}}.$$

The theorem follows using a similar argument as in the last part of the proof of Theorem 3 in Fernández and Steel (1998). □

Proof of Theorem 4

Again making use of Theorem 1 we prove the result using the symmetric sampling case. Having $E\left(|\Sigma|^{\frac{r}{2}}|Y\right) < \infty$ is equivalent to having

$$I = \int_{C^m \times \Re_+^n \times \mathcal{N}^*} |\Sigma|^r f_{iW}^m \left[\Sigma | S(\lambda), n-k\right] |\Lambda|^{\frac{m}{2}} |X'\Lambda X|^{-\frac{m}{2}} |S(\Lambda)|^{-\frac{n-k}{2}} d\Sigma d\lambda dP_{\lambda,\nu^*} < \infty \tag{28}$$

where $f_{iW}^m(\cdot|S, b)$ denotes the pdf of the inverted Wishart distribution with $m \times m$ scale matrix S and b df, and, thus, I can only be finite if $r < n - m - k + 1$. For such values of r, integrating Σ in equation (28) results in

$$I = \int_{\Re_+^n \times \mathcal{N}^*} |\Lambda|^{\frac{m}{2}} |X'\Lambda X|^{-\frac{m}{2}} |S(\lambda)|^{-\frac{n-k-r}{2}} dP_{\lambda,\nu^*}. \tag{29}$$

Following the procedure used in the last part of the proof of Theorem 3, and identifying $\lambda_{m_1}, \ldots, \lambda_{m_k}$ (Definition 3) and λ_b (Lemma 3) for each one of the $n!$ possible orderings of $\{\lambda_1, \ldots, \lambda_n\}$ provides an upper bound for the integrand in (29) proportional to

$$F(\lambda) = \lambda_b^{-\frac{m(n-k-r)}{2}} \prod_{i \neq m_1, \ldots, m_n} \lambda_i^{\frac{m}{2}}.$$

The proof is now concluded as in the proof of Theorem 4 (B) in Fernández and Steel (1998). □

Proof of Theorems 5, 6 and 7

Once again we use Theorem 1 and prove the results under symmetric sampling.

From the result in Lemma 6 we have that, if $\nu^\bullet$ equals the smallest component of ν, the df vector of a multivariate independent Student components distribution, and if $\nu^\bullet > m - 1$, then the pdf of this last distribution can always be bounded by a finite and bounded constant times the pdf of a m-dimensional Student pdf with $\nu^\bullet - m + 1$ df. Thus, using the result of Lemma 7, we can use the proofs of Theorems 2-4 to prove Theorems 5-7, respectively. □

B Orthogonal Matrices

In the sequel, O denotes an $m \times m$ orthogonal matrix with determinant equal to $(-1)^{m+1}$. The set of all such matrices will be denoted as O^m.

B.1 Householder matrices

Definition 4. Let v be a vector in $\Re^m$. Then,

$$H(v) = I_m - 2\frac{vv'}{v'v}$$

is a m-dimensional Householder matrix (e.g., Golub and van Loan, 1983, p. 38)). For any $v \in \Re^m$, $H(v)$ is an orthogonal, symmetric matrix of determinant -1 for which $H(v) = H(-v) = H(av)$, where a is a scalar. The latter property implies that if $v \in S_m^{1/2}$, a unit half-sphere in $\Re^m$, then v provides a one-to-one parameterisation of the set of m-dimensional Householder matrices.

Writing $v_\theta = (v_1, \ldots, v_m)'$ in polar coordinates, *i.e.*,

- $v_1 = \sin(\theta_1)$
- $v_j = \prod_{l=1}^{j-1} \cos(\theta_l) \times \sin(\theta_j), \quad j < m$
- $v_m = \prod_{l=1}^{m-1} \cos(\theta_l)$

and selecting $\theta = (\theta_1, \ldots, \theta_{m-1})'$ to be in Θ^m defined as,

- $(-\frac{\pi}{2}, \frac{\pi}{2})$ if $m = 2$
- $(0, \pi/2) \times (-\pi/2, \pi/2)^{m-3} \times (-\pi, \pi)$ if $m > 2$

implies that $v_\theta \in S_m^{1/2}$, and therefore by defining $H_\theta = H(v_\theta)$ we can uniquely parameterise the set of m-dimensional Householder matrices using $\theta \in \Theta^m$.

B.2 Decomposing O using Householder matrices

We now use Householder matrices to decompose any orthogonal matrix O, with $|O| = (-1)^{m+1}$. Let $\theta^j = (\theta_1^j, \ldots, \theta_{m-1}^j)' \in \Theta^j$ and for $j = 1, \ldots, m$ define $O_{\theta^j}^m$ as

$$O_{\theta^j}^m = \begin{pmatrix} I_{m-j} & 0 \\ 0 & H_{\theta^j} \end{pmatrix}. \tag{30}$$

Lemma 8. Any $m \times m$ orthogonal matrix O with determinant equal to $(-1)^{m+1}$ can be written uniquely as

$$O = O_{\theta^m}^m \times \ldots \times O_{\theta^2}^m, \tag{31}$$

where $\theta^j \in \Theta^j, \ j = 2, \ldots, m$.

Proof See Golub and Van Loan (1989), Chapter 5.

Thus, $O \in O^m$ can be parameterised uniquely by a set of $m-1$ vectors $\theta^j \in \Theta^j, \ j = 2, \ldots, m$.

B.3 Distribution on O^m invariant to linear orthogonal transformations

Stewart (1980) uses the decomposition in (31) to describe an algorithm for generating random orthogonal matrices from the invariant (with respect to linear orthogonal transformations) distribution on O^m (*i.e.* the Haar measure with respect to the orthogonal group). Using similar arguments, we can write the invariant distribution of O on O^m as

$$p(O) = p(\theta^2, ..., \theta^m) = \prod_{j=2}^{m} p(\theta^j), \tag{32}$$

where $p(\theta^j)$ is the pdf on $\theta^j \in \Theta^j$ that generates H_{θ^j} with first column uniformly distributed on $S_j,\ j = 2, \ldots, m$. Calculation reveals that the i-th element of the first column of H_{θ^j} is given by

- $\cos(2\theta_1^j)$ if $i = 1$
- $-\sin(2\theta_1^j)\prod_{l=2}^{i-1}\cos(\theta_l^j)\sin(\theta_i^j)$ if $2 < i \leq j-1$
- $-\sin(2\theta_1^j)\prod_{l=2}^{m-1}\cos(\theta_l^j)$ if $i = j$.

Variable transformation shows that, if

$$p(\theta^j) \propto |\sin(2\theta_1^j)^{j-2}\prod_{l=1}^{j-3}\cos(\theta_{l+1}^j)^{j-2-l}|,\ j = 2, \ldots, m, \tag{33}$$

then the first column of H_{θ^j} has an uniform distribution on the j-dimensional unit sphere S_j.

Equations (32)-(33) provide the necessary distribution of $\theta^j,\ j = 2, \ldots, m$, such that the distribution on O defined as in (31) is invariant on O^m.

B.4 Invariant distribution on $\mathcal{O}^m$

The invariant distribution on $\mathcal{O}^m$ is easily obtained as a restriction of the invariant distribution on O^m. If now, $O = (O_{ij}),\ i, j = 1, \ldots, m$, denotes an orthogonal matrix belonging to $\mathcal{O}^m$, then

- $O_{11} > -O_{m1} > -O_{(m-1)1} > \ldots > |O_{21}| > 0$
- $|O| = (-1)^{m+1}$.

Using $O = O_{\theta^m}^m \times \ldots \times O_{\theta^2}^m$, all the restrictions on O above are translated into restrictions on θ^m. Manipulation of the first column of H_{θ^m} shows that θ^m has to meet the following requirements:

$\mathbf{m = 2}: \theta^2 \in \left(-\frac{\pi}{8}, \frac{\pi}{8}\right)$

$$\mathbf{m > 2}: \theta^m \in \Bigg\{ \left(\theta_1^m, \ldots, \theta_{m-1}^m\right)' : \theta_{m-1}^m \in (a, \pi/4)\ \wedge\ \theta_j^m \in \left(0, \text{atan}\left[\sin\left(\theta_{j+1}^m\right)\right]\right),\ j = 2, \ldots, m-2\ \wedge$$
$$\wedge\ \theta_1^m \in \left(0, \frac{\text{acot}\left[\prod_{j=2}^{m-1}\cos\left(\theta_j^m\right)\right]}{2}\right)\Bigg\},\ a = -\frac{\pi}{4} \text{ if } m = 3,\ a = 0 \text{ otherwise.}$$

For $j = 2, \ldots, m-1,\ \theta^j \in \Theta^j$ as defined previously. As a consequence, the parameter space of $\theta^2, \ldots, \theta^m$ is always connected, which facilitates inference.

References

Arnold, B. C. and Beaver, R. J. (2000). Hidden truncation models, *Sankhyā A*, **62**: 23–35.

Arnold, B. C. and Beaver, R. J. (2002). Skewed multivariate models related to hidden truncation and/or selective reporting (with discussion), *Test*, **11**: 7–54.

Azzalini, A. (1985). A class of distributions which includes the normal ones, *Scandinavian Journal of Statistics*, **12**: 171–178.

Azzalini, A. and Capitanio, A. (1999). Statistical applications of the multivariate skew normal distribution, *Journal of the Royal Statistical Socienty*, Ser. B, **61**: 579–602.

Azzalini, A. and Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution, *Journal of the Royal Statistical Socienty*, Ser. B, **65**: 367–389.

Azzalini, A. and Dalla Valle, A. (1996). The multivariate skew-normal distribution, *Biometrika*, **83**: 715–726.

Bauwens, L. and Laurent, S. (2002). A new class of multivariate skew densities, with application to GARCH models, *Technical report*, CORE, University of Lovain.

Branco, M. and Dey, D. K. (2001). A general class of multivariate skew elliptical distributions, *Journal of Multivariate Analysis*, **79**: 99–113.

Cook, R. D. and Weisberg, S. (1994). *An Introduction to Regression Graphics*, Wiley, New York.

Cosh, A. D., Hughes, A., Keeble, D., Moore, B., Wilkinson, F. and Kitson, M. (2002). Cambridge Centre for Business Research SME Dataset, 1987-1995 [computer file]. 2nd ed. Colchester, Essex: UK Data Archive [distributor], 20 February 2002. SN: 4156.

Fang, K. T., Kotz, S. and Ng, K. W. (1990). *Symmetric Multivariate and Related Distributions*, Chapman and Hall, London.

Fernández, C. and Steel, M. F. J. (1998). On Bayesian modeling of fat tails and skewness, *Journal of the American Statistical Association*, **93**: 359–371.

Fernández, C. and Steel, M. F. J. (1999). Multivariate student-t regression models: Pitfalls and inference, *Biometrika*, **86**: 153–167.

Fernández, C. and Steel, M. F. J. (2000). Bayesian regression analysis with scale mixtures of normal, *Econometric Theory*, **16**: 80–101.

Ferreira, J. T. A. S. and Steel, M. F. J. (to appear). Bayesian multivariate skewed regression modelling with an application to firm size, *in* M. G. Genton (ed.), *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*, CRC Chapman & Hall.

Gantmacher, F. R. (1959). *The Theory of Matrices, vol. 1*, Chelsea, New York.

Gibrat, R. (1931). *Les inégalités économiques; applications: aux inégalités des richesses, à la concentration des enterprises, aux populations des villes, aux statistiques des familles, etc., d'une loi nouvelle, la loi de l'effet proportionnel*, Librairie du Recueil Sirey, Paris.

Golub, G. H. and van Loan, C. F. (1989). *Matrix Computations*, 2nd edn, John Hopkins University Press, Baltimore.

Hoggart, C. J., Walker, S. G. and Smith, A. F. M. (2003). Bivariate kurtotic distributions of garment fibre data, *Applied Statistics*, **52**: 323–335.

Jones, M. C. (2002). Marginal replacement in multivariate densities, with application to skewing spherically symmetric distributions, *Journal of Multivariate Analysis*, **81**: 85–99.

Jones, M. C. and Faddy, M. J. (2003). A skew extension of the t-distribution, with applications, *Journal of the Royal Statistical Society*, Ser. B, **65**: 159–174.

Kelker, D. (1970). Distribution theory of spherical distributions and a location-scale parameter generelization, *Sankhyā*, **32**: 419–430.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications, *Biometrika*, **57**: 519–530.

Moran, P. A. P. (1967). Testing for correlation between non-negative variates, *Biometrika*, **54**: 385–394.

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap (with discussion), *Journal of the Royal Statistical Socienty*, Ser. B, **56**: 3–48.

Sahu, S., Dey, D. K. and Branco, D. (2003). A new class of multivariate skew distributions with applications to Bayesian regression models, *The Canadian Journal of Statistics*, **31**: 129–150.

Stewart, G. W. (1980). The efficient generation of random orthogonal matrices with an application to condition estimators, *SIAM Journal on Numerical Analysis*, **17**: 403–409.

Sutton, J. (1997). Gibrat's legacy, *Journal of Economic Literature*, **35**: 40–59.